ECONSTOR
*Make Your Publications Visible.*

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Oberhofer, Harald; Pfaffermayr, Michael

**Article**

# Fractional Response Models - A Replication Exercise of Papke and Wooldridge (1996)

Contemporary Economics

**Provided in Cooperation with:**
University of Finance and Management, Warsaw

*Suggested Citation:* Oberhofer, Harald; Pfaffermayr, Michael (2012) : Fractional Response Models - A Replication Exercise of Papke and Wooldridge (1996), Contemporary Economics, ISSN 2084-0845, Vizja Press & IT, Warsaw, Vol. 6, Iss. 3, pp. 56-64, https://doi.org/10.5709/ce.1897-9254.50

This Version is available at:
https://hdl.handle.net/10419/105385

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Primary submission: 24.05.2012 | Final acceptance: 13.06.2012

# Fractional Response Models - A Replication Exercise of Papke and Wooldridge (1996)

Harald Oberhofer[1] and Michael Pfaffermayr[2,3]

**ABSTRACT** This paper replicates the estimates of a fractional response model for share data reported in the seminal paper of Leslie E. Papke and Jeffrey M. Wooldridge published in the Journal of Applied Econometrics 11(6), 1996, pp.619-632. We have been able to replicate all of the reported estimation results concerning the determinants of employee participation rates in 401(k) pension plans using the standard routines provided in Stata. As an alternative, we estimate a two-part model that is capable of coping with the excessive number of boundary values equalling one in the data. The estimated marginal effects are similar to those derived in the paper. A small-scale Monte Carlo simulation exercise suggests that the RESET tests proposed by Papke and Wooldridge in their robust form are useful for detecting neglected non-linearities in small samples.

**KEY WORDS:** replication exercise, fractional response models, two-part models, Monte Carlo simulation

**JEL Classification:** C15, C21

[1] University of Salzburg, Austria
[2] University of Innsbruck, Austria
[3] Austrian Institute of Economic Research, Austria

## Introduction

In many applications, the situation in which share data are confined to the [0,1] interval must be addressed, and in addition, the data may include a significant amount of observations of the dependent variable taking on values at the boundaries of 0 or 1. While share data can be handled using log-odds transformed variables, the combination of these two issues is complex. In their seminal paper, Leslie E. Papke and Jeffrey M. Wooldridge (1996) propose a fractional response model that extends the generalised linear model (GLM) literature from statistics. In a recent paper, Papke and Wooldridge (2008) introduce fractional response models for panel data. The authors introduce a quasi-maximum likelihood estimator (QLME) to obtain a robust method for estimating fractional response models without an *ad hoc* transformation of the boundary values. The paper shows that the proposed QLME is consistent given that the conditional mean function is correctly specified (see their equation 4). In addition, the authors introduce robust Ramsey RESET tests for the correct specification of the mean function. Finally, the paper provides an application of this estimation procedure: estimating a model of employee participation rates in 401(k) pensions plans. Ramalho, Ramalho and Murteira (2011) provide a comprehensive up-to-date overview on the econometrics of fractional response models.

Papke and Wooldridge (1996) consider the following model for the conditional expectation of the fractional response variable:

$$E(y_i \mid \mathbf{x}_i) = G(\mathbf{x}_i\boldsymbol{\beta}), i = 1,..., N, \quad (1)$$

where $0 \le y_i \le 1$ denotes the dependent variable and (the $1 \times k$ vector) $\mathbf{x}_i$ refers to the explanatory variables of observation $i$. Typically, $G(.)$ is a dis-

Corespondence concerning to this article should be addressed to: harald.oberhofer@sbg.ac.at

 DOI: 10.5709/ce.1897-9254.50

tribution function similar to the logistic function $G(z) = \exp(z)/(1+\exp(z))$, which maps $z$ to the $(0,1)$ interval. The authors follow the methods of McCullagh and Nelder (1991) and suggest maximising the Bernoulli log likelihood with the individual contribution given by the following (Papke & Wooldridge, 1993 also consider the case in which the group size is known and is given by $n_i$. They show that in this case, the conditional likelihood for observation $i$ is the same as that in (2), but it is weighted by $n_i$):

$$l_i(\boldsymbol{\beta}) = y_i \log[G(\mathbf{x}_i\boldsymbol{\beta})] + (1-y_i)\log[1-G(\mathbf{x}_i\boldsymbol{\beta})]. \qquad (2)$$

In this formulation of the likelihood function, the number of draws (here, the number of eligible employees of each firm) drops out be¶cause it does not depend on the parameters. Rather, the share of successes, i.e., the participation rate, enters the likelihood directly (see McCullagh & Nelder, 1991, p. 114).

The consistency of the QLME follows from the study by Gourieroux, Monfort & Trognon (1984) because the density upon which the likelihood function is based is a member of the linear exponential family, and the assumption that the conditional expectation of $y_i$ is correctly specified validates this finding. In fact, the QLME is $\sqrt{N}$ -asymptotically normal regardless of the distribution of $y_i$ conditional on $\mathbf{x}_i$. Papke and Wooldridge (1996) provide valid (robust) estimators of the asymptotic variance of $\boldsymbol{\beta}$ based on the well-known sandwich formula (see Cameron & Trivedi, 2005) and the non-linear conditional mean $G(.)$.

Papke and Wooldridge (1996) introduce and apply extended Ramsey RESET tests for $H_0: \gamma_1 = 0, \gamma_2 = 0$ in the augmented model $G(\mathbf{x}_i\boldsymbol{\beta} + \gamma_1(\mathbf{x}_i\boldsymbol{\beta})^2 + \gamma_2(\mathbf{x}_i\boldsymbol{\beta})^3)$. Their first RESET test is non-robust because it maintains the GLM variance assumption: $Var(y_i \mid \mathbf{x}_i) = \sigma^2 G(\mathbf{x}_i\boldsymbol{\beta})[1-G(\mathbf{x}_i\boldsymbol{\beta})]$. The robust RESET test only requires the correct specification of the conditional mean. Details on calculating the RESET test are provided on pages 623-625 in their paper.

In many applications, including that presented in this paper, there is a significant share of boundary values. Considering the data-generating process in the paper by Papke and Wooldridge (1996) literally, one would use the number of eligible employees as the number of Bernoulli draws. However, in the full sample, the mean firm size is 4621 and the median firm size is 628. Basing the Bernoulli draws on these numbers makes a boundary value of 1 in PRATE a rare event. Thus, in the case where 42.7 per cent of the boundary values in the data are equal to 1, it appears plausible to assume that firms that exhibit 100 per cent participation rates in their pension plans behave differently and are not well described by the Bernoulli model.

According to problem 19.8 in Wooldridge (2002), Ramalho and Vidigal da Silva (2009) and Ramalho, Ramalho and Murteira (2011), we can alternatively consider a two-part model that accounts for an excessive number of boundary values that are equal to one (refer to Pohlmeier and Ulrich (1995) for an early application of a two-part model for count data). We define:

$$y_i^* = \begin{cases} 0 & \text{if } y_i \,\varepsilon [0,1) \\ 1 & \text{if } y_i = 1 \end{cases} \qquad (3)$$

and assume for the first part of the model that $P(y_i^* = 1 \mid \mathbf{x}_i) = P(y_i = 1 \mid \mathbf{x}_i) = G(\mathbf{x}_i\gamma)$, where $G(\mathbf{x}_i\gamma)$ denotes the cumulative logistic distribution function. The second part of the model is the fractional response model that refers to observations $y_i \,\varepsilon\, [0,1)$. Then, the conditional mean of the two-part model is specified as the following:

$$E[y_i \mid \mathbf{x}_i] = P(y_i^* = 0 \mid \mathbf{x}_i)E[y_i \mid \mathbf{x}_i, y_i^* = 0] + P(y_i^* = 1 \mid \mathbf{x}_i)$$
$$= (1 - G(\mathbf{x}_i\gamma))G(\mathbf{x}_i\beta) + G(\mathbf{x}_i\gamma). \qquad (4)$$

The marginal effects of the explanatory variables can be derived as follows:

$$\frac{\partial E[y_i \mid \mathbf{x}_i]}{\partial x_{ij}} = \frac{\partial P(y_i^* = 1 \mid \mathbf{x}_i)}{\partial x_{ij}}(1 - E[y_i \mid \mathbf{x}_i, y_i^* = 0])$$
$$+ (1 - P(y_i^* = 1 \mid \mathbf{x}_i)\, \frac{\partial E[y_i \mid \mathbf{x}_i, y_i^* = 0]}{\partial x_{ij}}. \qquad (5)$$

This model allows the explanatory variables to affect the outcome ( $y_i = 1$ ) and size of $y_i$ at $y_i \,\varepsilon\, [0,1)$ in a different way. More importantly, the explanatory variables in the first and second parts of the model are not required to be the same. Under this specification (quasi) maximum likelihood estimation is straightforward because it can be separated into the estimation of the logit model explaining $P(y_i^* = 1 \mid \mathbf{x}_i)$ using all of the observations and the estimation of the parameters of the conditional density $f(y_i \mid \mathbf{x}_i, y_i^* = 0)$ based only on the observation where $y_i < 1$. Essentially, the conditional distribution

of $y_i \mid \mathbf{x}_i, y_i^* = 0$ is derived from the unconditional binomial distribution through division by $1-G(\mathbf{x}_i\boldsymbol{\beta})^{n_i}$ so that

$$f(y_i|\mathbf{x}_i, y_i^*=0)=\binom{n_i}{n_i y_i}G(\mathbf{x}_i\boldsymbol{\beta})^{n_i y_i}(1-G(\mathbf{x}_i\boldsymbol{\beta}))^{n_i(1-y_i)}(1-G(\mathbf{x}_i\boldsymbol{\beta})^{n_i})^{-1},$$

where $n_i$ is the number of eligible employees (see also Papke & Wooldridge, 1993). In the case where $n_i$ is large, the last term will be equal to approximately 1. In the following derivation, we neglect this term. In fact, the second part of the model is defined as the fractional response model, as introduced above. Again, the critical assumption that is necessary to obtain consistent parameters is the correct specification of the conditional mean, which now requires the correct specifications of $P(y_i^* = 1 \mid \mathbf{x}_i)$ and $E[y_i \mid \mathbf{x}_i, y_i^* = 0]$.

## The replication exercise

In their application, Papke and Wooldridge (1996) are interested in an econometric model of the participation rates in 401(k) pension plans. These plans are employer-sponsored pension plans in which employees are permitted to make pre-tax contributions and the employer may match part of the contribution. The dependent variable ($PRATE$) is defined as the number of active pension accounts divided by the number of employees eligible to participate for a sample of US manufacturing firms. The explanatory variables of their model include the plan match rate of the employer ($MRATE$), the log size of the firm measured in terms of employment and the square of this value, the plan's age and the square of this value and a dummy variable called $SOLE$ that indicates whether the 401(k) pension plan is the only plan of its type offered by the firm. In sum, the following specification is estimated in Tables II and III of this paper:

$$E(PRATE|\mathbf{x})=G(\beta_1+\beta_2 MRATE+\beta_3\log(EMP)+\beta_4\log(EMP)^2 + \beta_5 AGE+\beta_6 AGE^2+\beta_7 SOLE). \quad (6)$$

The linear specification assumes $G(z) = z$, while in the non-linear fractional response regression, G(.) is specified as a logistic function, i.e., $G(z) = \exp(z)/(1+\exp(z))$. In a second specification, the authors additionally include $MRATE^2$ as an explanatory variable.

Tables II and III in the paper report simple OLS estimates and the QMLE of the fractional response model. The estimates in Table II use only the observations where $MRATE < 1$, while the estimation results in Table III are based on all of the observations. There are no zeros in the dependent variable, but 42.7 per cent of the sample represents the firms in which all of the employees participate in 401(k) pension plans so that $PRATE = 1$.

In Table II of their paper, the authors report that the firm's matching rate has a significant positive impact. The log firm size and the age of the plan enter non-linearly. The impact of the log firm size is significantly negative, but increases for large firms. $AGE$ is significantly positive but also has a decreasing effect. Last, the variable $SOLE$ is insignificant.

In Table II of the paper, the OLS estimates are rejected by both the non-robust and robust RESET tests, suggesting that the linear model neglects important non-linearities. However, the signs of the estimated parameters are the same for the OLS and the QLME estimates for all variables. There is an important difference between the OLS and QMLE estimates because the RESET tests (both in their robust and non-robust versions) do not reject the fractional response model. Furthermore, the $R^2$ of the fractional response model is 6 percentage points higher compared to the linear model.

From an economic point of view, the difference between the two models is important because the fractional response model implies that there is a decreasing marginal effect of $MRATE$. The authors also conclude that simply adding ($MRATE^2$) to the linear model is not sufficient to capture this non-linearity. The results in Table III of their paper show that the basic results do not change if the models are estimated over the entire sample. The only clear differences are that the quadratic term in $MRATE$ is now significant and the RESET test does not reject the fractional response model that includes $MRATE^2$, but it rejects the baseline specification.

The authors estimated and tested the fractional response model using GAUSS-code. We were able to easily replicate and verify their estimated results using the 'now available' standard Stata code and specifically, the Stata procedure glm with options fam(bin), link(logit) and scale(x2) for non-robust standard errors and options fam(bin), link(logit) and rob for robust standard errors. In this way, we have been able to replicate each entry in Tables II and III. Therefore, the fractional response model is attractive because it can be easily estimated using standard econometric software. The Stata code is available upon request from the authors.

 DOI: 10.5709/ce.1897-9254.50

We also estimated the two-part model using the basic specification proposed by Papke and Wooldridge (1996), which is reported in the first two columns of Table II in their paper. As noted above, these estimates exclude observations where $MRATE > 1$. For comparison, we reproduced the corresponding estimates in Table 1. In the logit model of the two-part model, the same variables that enter the fractional response model determine whether all of the employees participate in the 401(k) pension plans. Approximately all of the explanatory variables are significant, and for $MRATE, \log(EMP)$ and $\log(EMP)^2$, we obtain the same signs as those in the fractional response model. In contrast to the results of the fractional response model, $AGE$ turns out to be insignificant, while $AGE^2$ is positive at a p-value slightly higher than $0.05$. The variable $SOLE$ is significantly positive, which is also in contrast to the estimate in the fractional response model.

**Table 1.** Results for the Restricted Sample

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Variable | OLS | QMLE | Two-Part Model | |
| | | | Logit | QMLE |
| *MRATE* | 0.156 | 1.390 | 1.504 | 0.895 |
| | (0.012) | (0.100) | (0.160) | (0.089) |
| | [0.011] | [0.107] | [0.166] | [0.097] |
| log(*EMP*) | -0.112 | -1.002 | -0.852 | -0.690 |
| | (0.014) | (0.111) | (0.200) | (0.092) |
| | [0.013] | [0.110] | [0.197] | [0.094] |
| log(*EMP*)$^2$ | 0.052 | 0.054 | 0.039 | 0.037 |
| | (0.001) | (0.007) | (0.013) | (0.006) |
| | [0.001] | [0.007] | [0.013] | [0.006] |
| *AGE* | 0.006 | 0.050 | -0.006 | 0.054 |
| | (0.001) | (0.009) | (0.014) | (0.007) |
| | [0.001] | [0.009] | [0.016] | [0.006] |
| *AGE*$^2$ | -0.000 | -0.001 | 0.001 | -0.001 |
| | (0.000) | (0.000) | (0.000) | (0.000) |
| | [0.000] | [0.000] | [0.000] | [0.000] |
| *SOLE* | -0.000 | 0.008 | 0.585 | -0.215 |
| | (0.006) | (0.047) | (0.078) | (0.039) |
| | [0.006] | [0.050] | [0.078] | [0.040] |
| *ONE* | 1.213 | 5.058 | 2.316 | 3.420 |
| | (0.051) | (0.427) | (0.740) | (0.354) |
| | [0.048] | [0.421] | [0.741] | [0.352] |
| Observations | 3,784 | 3,784 | 3,784 | 2,489 |
| SSR | 93.666 | 92.695 | - | 92.506 |
| SER | 0.157 | 0.438 | - | 0.390 |
| $R^2$ | 0.142 | 0.152 | - | 0.153 |
| RESET | 39.55 | 0.606 | - | 29.55 |
| | (0.000) | (0.738) | - | (0.000) |
| Robust-RESET | 45.36 | 0.782 | - | 23.85 |
| | (0.000) | (0.676) | - | (0.000) |

Notes: See Table II in Papke and Wooldridge (1996). In the logit model, the value of the dependent variable is one if all employees participate in the 401(k) pension plan and is zero otherwise. The QMLE of the two-part model is estimated only for *PRATE*<1.

The second part of the fractional response model uses observations where $PRATE < 1$. With the exception of the significant negative impact of *SOLE*, we obtain qualitatively similar results as those of Model 2 in Table II of Papke and Wooldridge (1996). However, in quantitative terms, the parameter estimates are quite different. The fit of the two-part model is comparable to the original estimates with $R^2$ amounting to $0.153$. Similar to the value of $R^2$ for the non-linear fractional response model in Papke and Wooldridge (1996), the value of $R^2$ for the two-part model is based on the predicted values of all of the observations, including the boundary values. However, both the robust and non-robust RESET tests are rejected, indicating a possible misspecification of the second part of the fractional response model.

The main advantage of both the fractional response model and the two-part model is their ability to capture non-linearities, particularly in the decreasing effect of the matching rate. Table 2 reproduces the marginal effects of the matching rate of the estimated model, which are presented in columns 1 and 2 in Table II of the paper. To obtain this result, *SOLE* is set equal to 0, $AGE = 13$ and $EMP = 200;4{,}620;100{,}000$. The partial effects are computed at the matching rate values of $0$, $0.5$ an $1$. While the marginal effect under the linear model amounts to 0.156, it diminishes for both the fractional response model and the two-part model. The fractional response model implies an increase in *PRATE* by 2.9 percentage points as a response to an increase in MRATE from 0 to 0.1. Under the two-part model, the effect is smaller and amounts to 2.1 percentage points. Conversely, at $MRATE = 1$, the marginal effect of the two-part model is 1.3 per cent compared to an effect of 1.2 per cent, which was implied by the fractional response model. Generally, the two-part model results in relatively small marginal effects at low values of *MRATE*, but in a less pronounced decrease in the marginal effects as *MRATE* increases.

When observing the in-sample predictions of the estimated models, we found two puzzling results. First, it can easily be observed from the specification of the conditional mean under the logistic link assumption, i.e., $G(z) = \exp(z)/(1+\exp(z))$, that both of the considered models rule out values of 1 in the dependent variable. Put differently, the models by definition always predict a value that is lower than one for those observations of *PRATE* that fall on the boundary 1.

Table 3 presents the calculations of the mean of the residuals resulting from the estimates in Table II in the paper as well as for the two-part model within each quintile of *PRATE* and, separately, for the values on the boundary cases where $PRATE = 1$. As expected, the residuals are positive for the values of $PRATE = 1$ for both the OLS estimation and the QMLE. Additionally, there is virtually no difference between the considered models.

**Table 2.** The Marginal Effects from the QMLE and the Two-Part Model

| | *EMP*=200 | | *EMP*=4,620 | | *EMP*=100,000 | |
|---|---|---|---|---|---|---|
| MRATE | QMLE | Two-Part | QMLE | Two-Part | QMLE | Two-Part |
| 0 | 0.172 | 0.164 | 0.288 | 0.214 | 0.273 | 0.195 |
| *0.5* | 0.100 | 0.113 | 0.197 | 0.176 | 0.182 | 0.157 |
| 1 | 0.054 | 0.063 | 0.118 | 0.127 | 0.106 | 0.115 |

Second, we observed systematic effects in the residuals of both the linear and non-linear models. For the observations where $PRATE < 1$, all of the considered models overpredict for the lower three quintiles of *PRATE* and underpredict for the two upper quintiles. The same pattern is found for the residuals of the two-part model. In fact, the residuals of the four estimated models in Table II of Papke and Wooldridge (1996) and those of the two-part model are highly correlated, with correlations as high as 0.99. As in many applications, there is only a minor difference between the linear and non-linear models in terms of the root mean squared prediction error, and using a logistic link function leads to only small improvements.

 DOI: 10.5709/ce.1897-9254.50

**Table 3.** Residuals from the OLS, the QMLE and the Two-Part Model

| | Prate | OLS | QMLE | Two-Part |
|---|---|---|---|---|
| 1st Quintile | 0.552 | -0.284 | -0.280 | -0.280 |
| 2nd Quintile | 0.720 | -0.111 | -0.111 | -0.111 |
| 3rd Quintile | 0.802 | -0.035 | -0.038 | -0.037 |
| 4th Quintile | 0.876 | 0.032 | 0.030 | 0.031 |
| 5th Quintile | 0.949 | 0.087 | 0.084 | 0.085 |
| *PRATE*=1 | 1.000 | 0.127 | 0.127 | 0.127 |
| Total | 0.937 | 0.000 | 0.000 | -0.000 |

Note: The figures are based on the means within the respective quintile.

## A small scale Monte Carlo exercise on the performance of the proposed RESET tests

To investigate the performance of the proposed RESET test, we established a small Monte-Carlo simulation exercise. Bernoulli random variables were generated using the predicted participation rates of column 4 of Table II in the paper, assuming that the reported parameters are the true values (see Equations 2 and 3). Because the Bernoulli random variable measures the number of successes in $n$ trials, we set $n = 10$ in the first experiment to generate a large share of ones (approximately 20 per cent). To obtain the share variables, we divided the resulting Bernoulli random number by $n$ (and similarly in the other experiments). The drawback of this design is that we obtained only nine different realisations of the generated random variable. In Experiment 2, we set $n = 1000$, while in Experiment 3, we allowed $n$ to vary and assumed that $n = EMP$. The latter experiment introduces additional heterogeneity and violates the nominal variance assumption because the log of the number of employees and its square are used as regressors (see equation 6 in the paper and the discussion below). Experiments 4 and 5 are identical to Experiments 2 and 3, but assume that the estimated logit model is the true data generating process for the boundary values. We generated a uniformly distributed random variable and set the simulated value of $PRATE$ to 1 if this random variable is lower than the predicted probability, as implied by the logit model. Then, we applied the two-part model and estimate a fractional response model using only the non-boundary values.

We calculated the bias and root mean squared error (RMSE) of the estimated parameters resulting from 10,000 replications of Monte Carlo experiments. Following the methods of Kelejian and Prucha (1999), we define the bias as $med(\widehat{\theta} - \theta)$ and RMSE as $(Bias^2 + (IQ/1.35)^2)^{0.5}$, where $IQ$ is the interquantile range. In all of the experiments, the estimated parameters are virtually unbiased. With the exception of Experiment 1, the RMSEs are relatively small. In particular, the RMSEs are considerably smaller than the standard errors reported in the paper, which originate from estimated models that have a significant share of boundary values.

To obtain the power curves of the RESET tests, we assume that $\gamma_1$ takes on values in the range of values including $\{-0.025,-0.015,-0.005,0,0.005,0.015,0.025\}$ and $\gamma_2$ is 1/5 th of $\gamma_1$. Because the power had a significantly low value in Experiment 1, we scaled the $\gamma$ - values for this experiment by a factor of 10. In each experiment, we added $\gamma_1(\mathbf{x}_i\boldsymbol{\beta})^2 + \gamma_2(\mathbf{x}_i\boldsymbol{\beta})^3$ to the linear predictor. Therefore, at $\gamma_1 = \gamma_2 = 0$, the share of rejections in the respective experiment is an estimate of the size of the RESET tests, and at $\gamma_1 \neq 0$ or $\gamma_2 \neq 0$, the value for the power of the test is obtained.

In Tables 4 and 5, the simulated size (in bold figures) and power of the RESET tests are displayed for a nominal size of 0.05. For each value of $g1$, the first lines in the tables refer to the non-robust RESET test and the second lines refer to the robust version. While the RESET tests are properly sized under Experiment 1 and 2, we find the correct size for only the robust RESET test under Experiment 3, as expected. Although the construction of the share variable often remains unobserved empirically, its calculation is important for estimating and testing the fractional response models, as argued by Papke and Wooldridge (1996). In

**Table 4.** Power and size of the RESET tests under the Fractional Response Model for Experiments 1, 2 and 3

| Experiment | | *g*2 | -0.050 | -0.030 | -0.010 | 0.000 | 0.010 | 0.030 | 0.050 |
|---|---|---|---|---|---|---|---|---|---|
| | *g*1 | | | | | | | | |
| 1 | -0.250 | | 1.000 | 1.000 | 0.939 | 0.250 | 0.993 | 0.549 | 0.292 |
| 1 | -0.250 | | 1.000 | 1.000 | 0.919 | 0.227 | 0.992 | 0.536 | 0.342 |
| 1 | -0.150 | | 1.000 | 0.363 | 1.000 | 0.989 | 0.668 | 0.103 | 0.791 |
| 1 | -0.150 | | 1.000 | 0.331 | 1.000 | 0.986 | 0.644 | 0.133 | 0.835 |
| 1 | -0.050 | | 0.999 | 1.000 | 0.754 | 0.192 | 0.049 | 0.565 | 0.965 |
| 1 | -0.050 | | 0.999 | 1.000 | 0.716 | 0.164 | 0.057 | 0.626 | 0.975 |
| 1 | 0.000 | | 1.000 | 0.995 | 0.229 | **0.046** | 0.152 | 0.780 | 0.984 |
| 1 | 0.000 | | 1.000 | 0.992 | 0.191 | **0.049** | 0.191 | 0.822 | 0.990 |
| 1 | 0.050 | | 1.000 | 0.801 | 0.053 | 0.121 | 0.378 | 0.890 | 0.993 |
| 1 | 0.050 | | 1.000 | 0.747 | 0.046 | 0.151 | 0.434 | 0.912 | 0.995 |
| 1 | 0.150 | | 0.813 | 0.082 | 0.277 | 0.545 | 0.779 | 0.969 | 0.997 |
| 1 | 0.150 | | 0.745 | 0.055 | 0.317 | 0.602 | 0.819 | 0.977 | 0.998 |
| 1 | 0.250 | | 0.120 | 0.225 | 0.676 | 0.836 | 0.924 | 0.987 | 0.998 |
| 1 | 0.250 | | 0.070 | 0.256 | 0.720 | 0.868 | 0.942 | 0.991 | 0.999 |
| | | *g*2 | -0.005 | -0.003 | -0.001 | 0.000 | 0.001 | 0.003 | 0.005 |
| | *g*1 | | | | | | | | |
| 2 | -0.025 | | 1.000 | 1.000 | 1.000 | 1.000 | 0.986 | 0.409 | 0.184 |
| 2 | -0.025 | | 1.000 | 1.000 | 1.000 | 1.000 | 0.986 | 0.406 | 0.183 |
| 2 | -0.015 | | 1.000 | 1.000 | 0.993 | 0.887 | 0.473 | 0.097 | 0.808 |
| 2 | -0.015 | | 1.000 | 1.000 | 0.993 | 0.884 | 0.467 | 0.098 | 0.807 |
| 2 | -0.005 | | 1.000 | 0.998 | 0.577 | 0.164 | 0.056 | 0.680 | 0.998 |
| 2 | -0.005 | | 1.000 | 0.998 | 0.567 | 0.158 | 0.054 | 0.679 | 0.998 |
| 2 | 0.000 | | 1.000 | 0.951 | 0.200 | **0.051** | 0.190 | 0.929 | 1.000 |
| 2 | 0.000 | | 1.000 | 0.947 | 0.195 | **0.052** | 0.193 | 0.928 | 1.000 |
| 2 | 0.005 | | 0.999 | 0.699 | 0.050 | 0.153 | 0.538 | 0.994 | 1.000 |
| 2 | 0.005 | | 0.999 | 0.689 | 0.050 | 0.154 | 0.541 | 0.993 | 1.000 |
| 2 | 0.015 | | 0.801 | 0.093 | 0.432 | 0.842 | 0.984 | 1.000 | 1.000 |
| 2 | 0.015 | | 0.791 | 0.085 | 0.435 | 0.844 | 0.984 | 1.000 | 1.000 |
| 2 | 0.025 | | 0.169 | 0.363 | 0.966 | 0.998 | 1.000 | 1.000 | 1.000 |
| 2 | 0.025 | | 0.160 | 0.362 | 0.966 | 0.998 | 1.000 | 1.000 | 1.000 |
| 3 | -0.025 | | 1.000 | 1.000 | 0.985 | 0.925 | 0.753 | 0.264 | 0.200 |
| 3 | -0.025 | | 1.000 | 1.000 | 0.983 | 0.916 | 0.729 | 0.209 | 0.083 |
| 3 | -0.015 | | 1.000 | 0.988 | 0.790 | 0.534 | 0.291 | 0.160 | 0.454 |
| 3 | -0.015 | | 1.000 | 0.985 | 0.753 | 0.472 | 0.209 | 0.059 | 0.312 |
| 3 | -0.005 | | 0.991 | 0.814 | 0.345 | 0.181 | 0.146 | 0.380 | 0.814 |
| 3 | -0.005 | | 0.987 | 0.769 | 0.234 | 0.087 | 0.051 | 0.263 | 0.762 |
| 3 | 0.000 | | 0.953 | 0.597 | 0.200 | **0.138** | 0.197 | 0.586 | 0.927 |
| 3 | 0.000 | | 0.935 | 0.499 | 0.093 | **0.050** | 0.098 | 0.501 | 0.910 |
| 3 | 0.005 | | 0.851 | 0.390 | 0.140 | 0.179 | 0.333 | 0.779 | 0.977 |
| 3 | 0.005 | | 0.792 | 0.250 | 0.049 | 0.092 | 0.242 | 0.740 | 0.975 |
| 3 | 0.015 | | 0.449 | 0.158 | 0.286 | 0.494 | 0.721 | 0.973 | 0.999 |
| 3 | 0.015 | | 0.283 | 0.053 | 0.215 | 0.440 | 0.701 | 0.970 | 0.999 |
| 3 | 0.025 | | 0.190 | 0.254 | 0.687 | 0.862 | 0.962 | 0.997 | 1.000 |
| 3 | 0.025 | | 0.064 | 0.199 | 0.671 | 0.856 | 0.962 | 0.998 | 1.000 |

Notes: The DGP is assumed to be Model 4 reported in Table II of Papke and Wooldridge (1996). Bold figures refer to the size of the test; the remaining figures refer to the power. For each value of $g1$, the first line in the table refers to the non-robust version of the RESET test and the second line refers to the robust version.

Experiment 1: Bernoulli random variable scaled by 10.

Experiment 2: Bernoulli random variable scaled by 1000.

Experiment 3: Bernoulli random variable scaled by employment.

 DOI: 10.5709/ce.1897-9254.50

**Table 5.** Power and size of the RESET tests under the Two-Part Model for Experiments 4

| Experiment | | *g*2 | -0.005 | -0.003 | -0.001 | 0.000 | 0.001 | 0.003 | 0.005 |
|---|---|---|---|---|---|---|---|---|---|
| | *g*1 | | | | | | | | |
| 4 | -0.025 | | 1.000 | 1.000 | 0.997 | 0.977 | 0.862 | 0.269 | 0.093 |
| 4 | -0.025 | | 1.000 | 1.000 | 0.997 | 0.977 | 0.857 | 0.264 | 0.091 |
| 4 | -0.015 | | 1.000 | 0.999 | 0.879 | 0.609 | 0.290 | 0.065 | 0.434 |
| 4 | -0.015 | | 1.000 | 0.999 | 0.871 | 0.600 | 0.280 | 0.062 | 0.431 |
| 4 | -0.005 | | 0.999 | 0.904 | 0.328 | 0.104 | 0.053 | 0.354 | 0.895 |
| 4 | -0.005 | | 0.999 | 0.897 | 0.315 | 0.097 | 0.054 | 0.354 | 0.893 |
| 4 | 0.000 | | 0.988 | 0.676 | 0.119 | **0.051** | 0.115 | 0.631 | 0.974 |
| 4 | 0.000 | | 0.987 | 0.660 | 0.112 | **0.049** | 0.117 | 0.628 | 0.973 |
| 4 | 0.005 | | 0.922 | 0.375 | 0.051 | 0.106 | 0.300 | 0.859 | 0.996 |
| 4 | 0.005 | | 0.914 | 0.358 | 0.050 | 0.105 | 0.298 | 0.859 | 0.996 |
| 4 | 0.015 | | 0.430 | 0.069 | 0.265 | 0.553 | 0.826 | 0.994 | 1.000 |
| 4 | 0.015 | | 0.410 | 0.064 | 0.263 | 0.555 | 0.823 | 0.994 | 1.000 |
| 4 | 0.025 | | 0.091 | 0.245 | 0.787 | 0.943 | 0.991 | 1.000 | 1.000 |
| 4 | 0.025 | | 0.086 | 0.246 | 0.788 | 0.943 | 0.991 | 1.000 | 1.000 |
| 5 | -0.025 | | 1.000 | 0.993 | 0.892 | 0.771 | 0.588 | 0.259 | 0.162 |
| 5 | -0.025 | | 0.999 | 0.986 | 0.854 | 0.706 | 0.492 | 0.152 | 0.060 |
| 5 | -0.015 | | 0.990 | 0.902 | 0.601 | 0.411 | 0.270 | 0.153 | 0.274 |
| 5 | -0.015 | | 0.982 | 0.854 | 0.481 | 0.290 | 0.149 | 0.051 | 0.150 |
| 5 | -0.005 | | 0.907 | 0.624 | 0.284 | 0.188 | 0.153 | 0.243 | 0.546 |
| 5 | -0.005 | | 0.851 | 0.485 | 0.137 | 0.070 | 0.047 | 0.143 | 0.453 |
| 5 | 0.000 | | 0.791 | 0.456 | 0.205 | **0.154** | 0.164 | 0.361 | 0.689 |
| 5 | 0.000 | | 0.683 | 0.285 | 0.073 | **0.053** | 0.067 | 0.268 | 0.623 |
| 5 | 0.005 | | 0.647 | 0.316 | 0.160 | 0.164 | 0.239 | 0.517 | 0.822 |
| 5 | 0.005 | | 0.481 | 0.147 | 0.047 | 0.068 | 0.141 | 0.438 | 0.788 |
| 5 | 0.015 | | 0.344 | 0.169 | 0.222 | 0.334 | 0.486 | 0.804 | 0.961 |
| 5 | 0.015 | | 0.151 | 0.052 | 0.146 | 0.269 | 0.432 | 0.780 | 0.954 |
| 5 | 0.025 | | 0.201 | 0.225 | 0.473 | 0.636 | 0.785 | 0.956 | 0.995 |
| 5 | 0.025 | | 0.061 | 0.150 | 0.436 | 0.614 | 0.767 | 0.956 | 0.995 |

Notes: The DGP is assumed to be Model 4 reported in Table II of Papke and Wooldridge (1996). Bold figures refer to the size of the test; the remaining figures refer to the power. For each value of *g*1, the first line in the table refers to the non-robust version of the RESET test and the second line refers to the robust version.
Experiment 4: Bernoulli random variable scaled by 1000. Logit model of Table 1 is assumed to be the DGP.
Experiment 5: Bernoulli random variable scaled by employment. Logit model of Table is assumed to be the DGP.

this respect, our findings confirm the discussion of the RESET tests in the paper. The results of Experiments 4 and 5 refer to the two-part model and confirm the findings of Experiments 2 and 3. Alternatively, we also investigated the case in which a fractional response model using all of the observations is estimated in the case of a large share of boundary values. The results, which are available upon request from the authors, indicate that in this case, the RESET tests are oversized and their power tends to be considerably lower, even when taking into account that the tests are oversized. However, this finding has to be expected because this setup violates the conditional mean assumption.

Generally, the RESET tests exhibit sufficient power to detect neglected non-linearities. Only at small $n$ values, as in Experiment 1, the power is not satisfactory. For this experiment, we obtained power figures comparable to the other experiments when scaling $\gamma_1$ and $\gamma_2$ by a factor of 10. The highest power of the RESET test is observed when either $\gamma_1$ or $\gamma_2$ is zero and the corresponding non-zero value has a high absolute value. However, at large absolute values of $\gamma_1$ and $\gamma_2$ with different signs, the power of the RESET test results in a very low value. This result holds for the robust and non-robust versions of the RESET test.

## Conclusions

This paper replicated the results of the seminal paper of Leslie E. Papke and Jeffrey M. Wooldridge (1996) concerning a fractional response model for employee participation rates in 401(k) pension plans in US manufacturing firms. Using the 'now available' standard Stata code, we have been able to replicate each estimation result in the paper.

An important feature of their dependent variable is that more than 40 per cent of these data are equal to one, indicating full employee participation. To cope with the excessive number of boundary values, we additionally estimated a two-part model. The first part of this model estimates the probability of a boundary observation by a simple logit model. The second part of the model refers to non-boundary values and is estimated by the same fractional response model. The estimation of the second part of the model yields slightly different results. However, the marginal effects of the matching rate that take both parts into account are comparable in size. The effects are slightly smaller, and the diminishing impact of the matching rate is less pronounced. Therefore, in the presence of a high share of boundary values, the two-part model is a useful alternative to the fractional response model. Moreover, it is as easy to perform the calculations using this model with the available standard software.

Looking at the in-sample predictions of the estimated model reveals some complexities. First, for all of the observations with a boundary value of one in the dependent variable, the corresponding predictions by definition are less than one. Second, in all of the estimated models, there are systematic differences in the remaining residuals, depending on the size of the participation rate. A small-scale Monte Carlo simulation exercise confirms that the proposed RESET tests are useful for detecting neglected non-linearities in small samples. In their robust form, the RESET tests are always properly sized and equipped with power in approximately all of the considered cases.

## References

Cameron, C. A., & Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications.* Cambridge: Cambridge University Press.

Gourieroux, C., Monfort, A, & Trognon, A. (1984). Pseudo-maximum likelihood methods: Theory. *Econometrica*, 52(3), 681-700.

McCullagh, P., & Nelder, J. A. (1991). *Generalized Linear Models*. (2nd ed.), London, UK: Chapman and Hall.

Kelejian, H. H., & Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. I*nternational Economic Review*, 40(2), 509-533.

Papke, L. E., & Wooldridge, J. M. (1993). *Econometric methods for fractional response variables with an application to 401(k) plan participation rates* (Technical Working Paper No. 147). National Bureau of Economic Research.

Papke, L. E, & Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11(6), 619-632.

Papke, L. E., & Wooldridge, J. M. (2008). Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics*, 145(1-2), 121-133.

Pohlmeier, W., & Ulrich, V. (1995). An econometric model of the two-part decisionmaking process in the demand for health care. *Journal of Human Resources*, 30(2), 339-361.

Ramalho, J. J. S., & Vidigal da Silva, J. (2009). A two-part fractional regression model for the financial leverage decisions of micro, small, medium and large firms. *Quantitative Finance*, 9(5), 621-636.

Ramalho, E. A., Ramalho, J. J. S., & Murteira, J. M. R. (2011). **Alternative estimating and testing em**pirical strategies for fractional regression models. *Journal of Economic Surveys*, 25(1), 19-68.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

 DOI: 10.5709/ce.1897-9254.50